INTERNATIONAL JOURNAL OF ADVANCED STUDIES IN COMPUTER SCIENCE AND ENGINEERING
IJASCSE VOLUME 5, ISSUE 5, 2016
05/31/2016

Investigation of Thinning Techniques for Arabic Document Binary Images Acquired via Handheld Cameras

Musab Kasim Alqudah[1], Mohammad F. Nasrudin[2]
Pattern Recognition Research Group, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600, Bangi, Selangor, Malaysia

Arwa Mahmoud Alkhatatneh
Department of Computer Science, Faculty of Science and Technology
Universiti Sains Islam Malaysia (USIM)
71800, Nilai, Negeri Sembilan, Malaysia

***Abstract*— A mobile phone camera is one of the most flexible, easiest and fastest method for converting traditional to digital document images. However, the conventional binarization problems of document captured by handheld cameras is rarely investigated. Notably, all subsequent stages of document images analysis and recognition (DIAR) are influenced by the result of pre-processing stage in which the thinning technique of document pre-processing stage is employed to extract the text topology using in the feature extraction and recognition stages. This technique is a critical step to recognize the scripts, font, characters and symbols marks in various document applications. The objectives of this survey to explore the best thinning method based on Arabic document images, and investigate the effect of thinning technique on Arabic diacritics marks. The experiments are conducted on 20 Arabic document images based on Zhang-Susen, Huang, K3M and Abu-Ain thinning techniques. The results revealed that Abu-Ain technique outperforms the other techniques in terms of extracting the minimal pixels values.**

Keywords- Thinning; Binarization; Phone camera; Arabic font

I. INTRODUCTION *(HEADING 1)*

The pattern recognition is part and parcel of machine learning as it is involves several step to recognize the image elements (e.g., the DIAR applications (Marinai 2008; Kasturi et al. 2012)). DIAR applications are divided into two branches: textual applications are included printed, calligraphy "manuscript" and handwritten documents, and graphical applications were included pictures, figures and graphics. Each branch consists of two main stages: processing and recognizing of document. The both stages of each application can share same techniques to process the document or recognize the elements of objects, text or graphics (Marinai 2008). A phone or handheld camera is used to acquire document images from original or hardcopy document images, wherein the challenges in document images arise i.e., poor quality, skew and uneven contrast (Alqudah et al. 2015; Chou et al. 2010). These type of problems can exert influence in DIAR stages such as the binarization methods in pre-processing stage (Bataineh et al. 2011; Al-khatatneh et al. 2015).

Thinning technique is one of best method in pre-processing to retrieve the topology of object for searching the features of document elements, by means of minimizing the body width of the objects or text (Chen et al. 2012; Abu-Ain et al. 2013). The thinning or skeleton method is of great importance for wide application in enhancing the feature extraction techniques by extracting text skeleton in preprocessing stage of DIAR applications in binary images. Myriads of document analysis applications are based on thinning techniques as Optical Font Recognition (OFR) (Tellache et al. 1993; Lutf et al. 2014), Optical Character Recognition (OCR) (Lorigo and Govindaraju 2006; Ali 2012) and OSR (Ghosh et al. 2010).

The aims of thinning technique is to remove superfluous information and reduce the text to minimum pixels in order to obtain the text topology, which can have an impact on the accuracy of the results in pre-processing as well the recognition stages. There are two approaches of thinning methods available as shown in Fig.1: the iterative approach and the non-iterative approach that exploits the behavior of pixels (Huang et al. 2003; Saeed et al. 2010; Nemeth and Palagyi 2011; Abu-Ain et al. 2013).

The non-iterative techniques, which are based on elements edge, extract the skeleton of objects from standard or non-standard polygons (Chu et al. 2013). Whereas, the iterative approach is divided into parallel (e.g., the work of Huang (2003)) and sequential methods (e.g., the work of Saeed et al. (2010)).

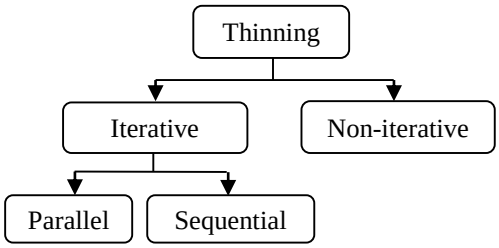


Figure 1. The main thinning approaches

The performance of the thinning method hinges on the condition and properties of images, whereby the condition of images is a pressing factor in dealing with objects and text as evidenced by the fact that the binarization result is introduced in thinning step of document images analysis. In addition, the properties of objects and text require different steps to extract the skeleton based on shape of objects, cursive text and additional symbol marks (e.g., Arabic decorations or diacritics), whereby problems such as in text dis-connectivity, inconsistent topologies and omission of the tails pixels or symbols encumbers the thinning process. Fig. 2 illustrates the general steps of thinning techniques applied to extract the skeleton topology, in which the following hierarchical flow chart identifies the general framework of thinning.

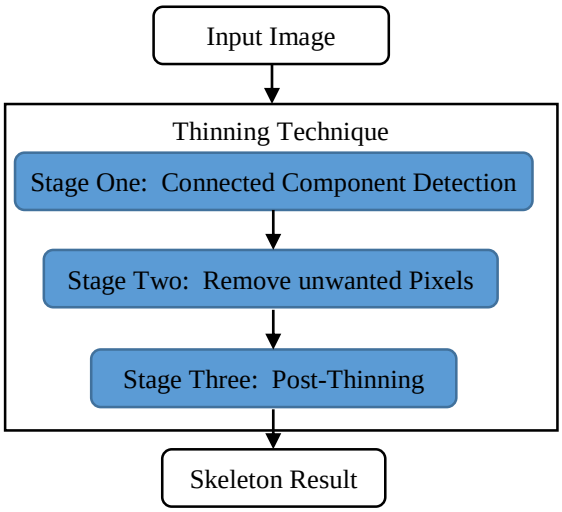


Figure 2. The General framework of thinning Technique.

This paper presents the performance evaluation for the fourth thinning methods (Zhang-Susen, Huang, K3M and Abu-Ain) based on fine Arabic document images captured via phone camera and the thinning results are evaluated against the ground truth of binarization method. The twenty Arabic printed document images consists of the Arabic decoration and non-decoration symbol marks. We presented four thinning methods in Section 2, the experiments in Section 3, and finally the conclusion in Section 4.

II. State of the Art

In the state of the art section, we presented the four well-known thinning methods applied on text and characters. The iterative approach was often regarded as a good technique due to its dependence on process of iterative edge removal to acquire the best skeleton form (one pixel width). The parallel and sequential techniques were utilized to remove the pixels based on their requirement, in which sequential was employed to identify the required pixels and thereby, eliminate non-required pixels, while parallel identified the non-required pixels and thereby, processing by eliminating these pixels (Zhang and Suen 1984; Huang et al. 2003; Saeed et al. 2010; Abo-Ain et al. 2013). Zhang-Susen (Zhang and Suen 1984) and Huang (Huang et al. 2003) represented the parallel-thinning technique of iterative approach, and K3M (Saeed et al. 2010) and Abu-Ain (Abo-Ain et al. 2013) represented sequential technique of iterative approach. Brief overviews of the methods are described in the following subsections.

A. *Zhang-Susen*

Zhang and Suen method (1984) was considered the primary parallel methods in iterative approach (Chen et al. 2012). The Zhang-Suen method utilized on 3*3 mask of thinning eleven template to eliminate the unwanted duplication pixels in each round. This method retained good connectivity of objects or text. This method was easily affected by thick dark pixels, i.e., cursive language as Arabic words, which in return retrieve two pixels width of skeleton results.

B. *Huang*

Huang method (2003) enhanced the parallel technique by improving the removal rules and different mask size, which optimized reduction of skeleton width. This method rules were based on the relationship between neighboring pixels in each mask window size, and were exercised concurrently in weight values of pixels while applying other rules to remove the additional pixels in connected neighbor pixels. However, all of these rules were based on mask window size (e.g., 4*4, 4*3, and 3*4). In addition, this method was based on thresholding to minimize the pixels in some cases of arbitrary thinning shape. This method could be negatively impacted by pixel loss and test dis-connectivity.

C. *K3M*

Saeed et al. (2010) enhanced the KMM method (Saeed & Niedzielski 1999; Saeed 2001) was oriented towards the sequential method, known as K3M. This method was based on seven stages of determining and removing unwanted pixels. This method involved the relationship between neighboring pixels with weights of mask window. The seven stages were executed in three phases: first phase of initial stage was established to determine the edge of connected components, the second phase through next fifth stages were ensued to remove the pixels, and the final phase in stage seven employed differential equations to modify and enhance the

distorted pixels. However, this method did not take into account the time and complexity in implementation, causing the performance to deteriorate due to large image processing.

D. Abu-Ain

Abu-Ain et al. (2013) proposed sequential method was conducted in iterative approach, entailing on three phases to extract the skeleton body. In phase one, the contour detection was performed to identify the element edges of connected component. In phase two, the processing and removal pixels employed 3*3 mask to unwanted element pixels based on relationship between pixels and connected component connectivity. The last phase reused 3*3 mask window to eliminate the duplicate pixels of connected component. This method overcame other methods in several problem cases (e.g., superior tails, topology of elements and extract minimum pixel width).

III. EXPERIMENT

The experimental results was based on twenty Arabic document images captured from a 5MP phone camera and processed by binarization technique; eight of them were Arabic document images containing diacritics symbol marks "Harakat" and the remainder document images without Harakat. The diacritics Arabic symbol signified "Harakat" depending on Tashkeel style (e.g.,گ ,ً, ٌ ,~). This small symbols located above or below each character were omitted by most thinning techniques, i.e., vowel marks, but utilized in the next stages for recognition in the various text in document processing and analysis applications.

These Arabic document images were set as benchmark dataset in our faculty and part of these dataset have been made available under our research center website: http://www.ftsm.ukm.my/cait/index.php/download/category/8-dataset (Alqudah et al. 2015).

The visual experiments illustrated the skeleton results of selected images for proposed methods as shown in Fig.3. The challenges of thinning techniques that remained unsolved were rotation angles in circle pixels, the width of pixels, superior tails of connected component, symbols distortions and topological distortions and connectivity. The results of Abu-Ain method showed significant advantage over the other method in some challenges (e.g., the thin and rotation of connected components). Nonetheless, this skeleton method had yet to make the grade.

IV. CONCLUSION

The thinning methods were shown to be imperative in extracting the topology of document contains. This is because the skeleton of document results is adopted in various document processing and analysis applications in order to classify and recognize the image data. Therein, numerous techniques are employed to determine the relationship occurred between pixels connectivity. The aims of thinning techniques are to preserve significant image topologies, to sufficiently accommodate the details of pattern types of images, the connectivity of connected component, while generating one pixels width and keeping the end points of image connected (e.g., objects and text).

The experiment conducted on Arabic textual document support the hypothesis of the study that the thinning approach is not efficient for retaining of small connected component in images, albeit suitable for preserving the topology or geometric features of document images. Also, the visual experiments demonstrated various challenges that appear in the results of selected methods such as non-uniform of pixel width, different retrieval to maintain topology and uneven generation of tail.

REFERENCES

[1] W. Abu-Ain, S. N. H. S. Abdullah, B. Bataineh, T. Abu-Ain, and Omar, K., "Skeletonization Algorithm for Binary Images", Procedia Technology, 11, pp.704-709, 2013.

[2] M. A. Ali, "An efficient thinning algorithm for Arabic OCR systems", An International Journal Signal & Image Processing (SIPIJ). 3(3): 31-38, 2012. .

[3] A. M. Al-khatatneh, S. A. Pitchay, and M. K. Al-qudah, "Compound Binarization for Degraded Document Images," ARPN Journal of Engineering and Applied Sciences. ISSN 1819-6608. VOL. 10, NO. 2, February 2015, pp. 594-599.

[4] M. K. Alqudah, M. F. Nasrudin, B. Bataineh, M. Alqudah, and A. Alkhatatneh, "Investigation of binarization techniques for unevenly illuminated document images acquired via handheld cameras," IEEE. In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on Apr 21, 2015, pp. 524-529.

[5] B. Bataineh, S. N. H. S. Abdullah, and K., Omar, "An adaptive local binarization method for document images based on a novel thresholding method and dynamic windows," Pattern Recognition Letters, 2011, 32: pp. 1805-1813.

[6] H. Chou, H. Lin, and F. Chang, "A Binarisation Method with Learning-Built Rules for Document Images Produced By Cameras," Pattern Recognition, 2010, 43(4): pp. 1518-1530.

[7] W. Chen, L. Sui, Z. Xu, and Y. Lang, "Improved Zhang-Suen thinning algorithm in binary line drawing applications," In Systems and Informatics (ICSAI), IEEE. 2012 International Conference on May 2012, pp. 1947-1950.

[8] D. Ghosh, T. Dube, and A. P. Shivaprasad, "Script recognition - A review," IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI, 2010, 3212: pp. 2142-2161.

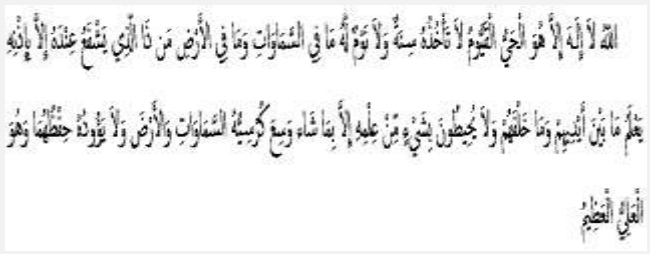
[9] L. Huang, G. Wan, and C. Liu, "An improved parallel thinning algorithm", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).780–783 .Pattern Analysis and Machine Intelligence, 2003: 14: pp. 869-885.

[10] M. Lutf, X. You, Y. M. Cheung, and C. P. Chen, "Arabic font recognition based on diacritics features," Pattern Recognition, 2014, 472: pp. 672-684.

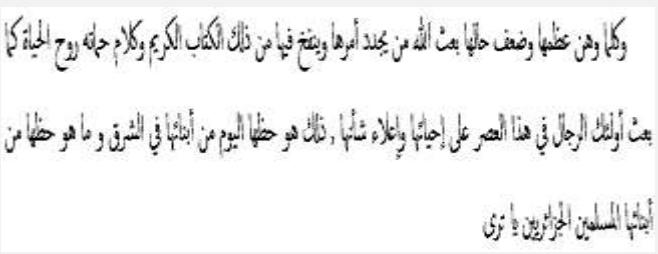
[11] G. Nemeth, and K. Palagyi, "Topology preserving parallel thinning algorithm," International Journal of Imaging System and Technology, 2011, No. 21, pp. 37-44.

[12] Saeed, K., Tabedzki, M., Rybnik, M., Adamski, M., K3M: A universal algorithm for image skeletonization and a review of thinning techniques. Applied Mathematics and Computer Science, 2010: 20, 317-335.

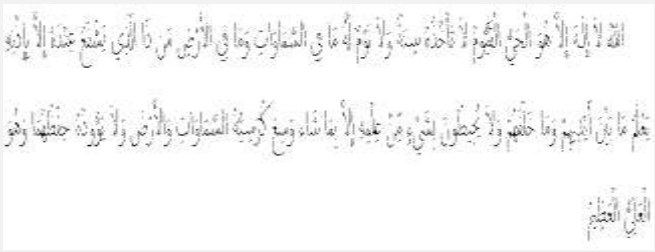
[13] T. Y. Zhang, and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," Communications of the ACM. 1984, No. 273, pp. 236-239.



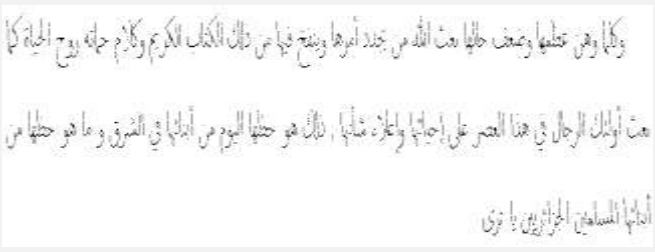
(a) First sample of binary Arabic document images



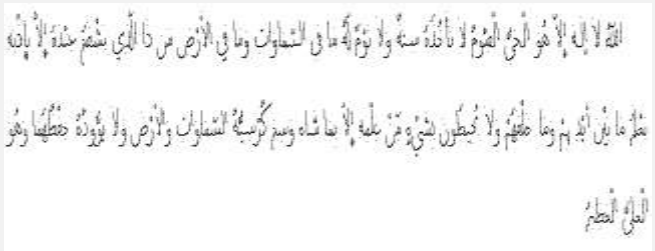
(a) Secound sample of binary Arabic document images



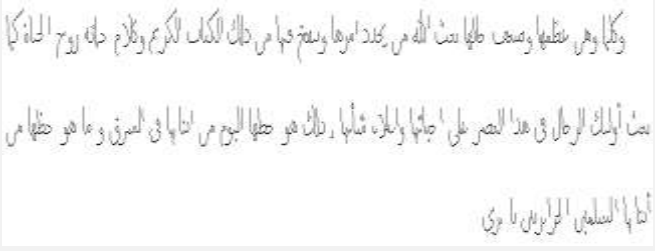
(b) Result of Zhang-Susen technique



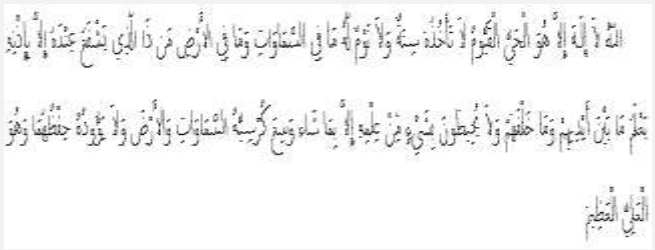
(b) Result of Zhang-Susen technique



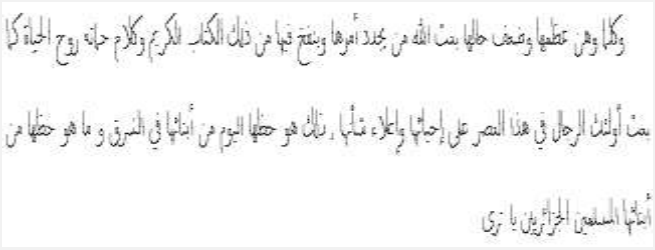
(c) Result of Huang technique



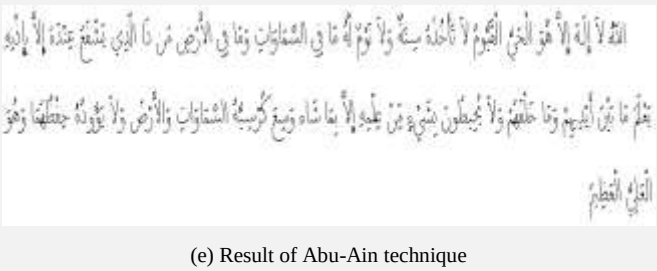
(c) Result of Huang technique



(d) Result of K3M technique



(d) Result of K3M technique



(e) Result of Abu-Ain technique

(e) Result of Abu-Ain technique

Figure 3. Samples of Arabic document image, one contains Tashkeel and other one without Tashkeel. (a) The original images. (b) Zhang-Suen results. (c) Huang results. (d) K3M results. (e) Abu-Ain results.